Charles University

Faculty of Science

Study program: Bioinformatics

Branch of study: Bioinformatics

Alžběta Šrůtková

Study of DNA hydration by analysis of structural data from databases

Studie hydratace DNA analýzou strukturních dat z databází

Bachelor's thesis

Supervisor: prof. Ing. Bohdan Schneider, CSc.

Prague, 2020

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne 6.června 2020 Podpis autora

Děkuji svému školiteli prof. Ing. Bohdanu Schneiderovi, CSc., za možnost vypracování bakalářske práce, odborné konzultace a trpělivou pomoc při jejím psaní. Dále děkuji Ing. Michaele Nekardové, Ph.D. a RNDr. Ladě Biedermnnové, Ph.D. za poskytnutí výsledků a vstřícné vedení v průběhu projektu. Děkuji také všem dalším členům týmu za pomoc a rady při praktické části práce.

Abstract

The deoxyribonucleic acids play an essential role in the living organisms by storing the genetic information. The process of the genetic information transfer, translation and transcription is carried out by other molecules, mostly proteins, which cause DNA bend and locally deform. These protein-DNA interactions, together with the nucleotide sequence, result in different DNA conformations. The description and classification of the DNA local structural diversity at a level of dinucleotides is provided by so-called dinucleotide conformer classes, NtC. The NtC classification system developed for description of the local conformations of nucleic acids consists of 97 NtC classes. In this work we use 44 classes important for the description of DNA conformations.

All biomolecules, including DNA are heavily hydrated. The first hydration layer around DNA is largely localized. The structure of this hydration layer depends on the sequence and conformation of the analyzed biomolecule. The DNA structural classification by the NtC classes allows specific study of the first hydration layer with a sufficient sequence and structural granularity. Based on the data available in the structural databases, we studied hydration of double helical tetranucleotide fragments in the three biologically most important DNA conformational classes: BI form (BB00 NtC class), A form (AA00 NtC class), and BII form (BB07 NtC class).

Key Words: DNA hydration, DNA structure, DNA conformation, structure database, NDB, PDB

Abstrakt

Deoxyribonukleové kyseliny hrají zásadní roli v živých organismech tím, že ukládají genetickou informaci. Proces přenosu, translace a transkripce genetické informace je prováděn molekulami, většinou proteiny, které způsobují ohyb a lokální deformaci vázané DNA. Tyto interakce spolu s nukleotidovou sekvencí formují DNA do různých konformací. Popis a klasifikace lokální strukturní diverzity DNA na úrovni dinukleotidů je poskytován tzv. třídami dinukleotidových konformerů, NtC. Klasifikační systém NtC vytvořený pro popis lokálních konformací nukleových kyselin se skládá z 97 tříd NtC. V této práci diskutujeme 44 tříd důležitých pro popis DNA konformací.

Všechny biomolekuly, včetně DNA, jsou silně hydratované. První hydratační vrstva kolem DNA je z velké části lokalizována. Struktura této hydratační vrstvy závisí na sekvenci a konformaci analyzované biomolekuly. Strukturální klasifikace DNA do tříd NtC umožňuje specifické studium první hydratační vrstvy s dostatečnou sekvenční a strukturní granularitou. Na základě údajů dostupných ve strukturních databázích jsme studovali hydrataci dvoušroubovicových tetranukleotidových fragmentů ve třech biologicky nejdůležitějších konformačních třídách DNA: BI forma (NtC třída BB00), A forma (NtC třída AA00) a BII forma (NtC třída BB07).

Klíčová slova: DNA hydratace, struktura DNA, konformace DNA, strukturní databáze, NDB, PDB

Abbreviations

NA	Nucleic Acid
DNA	Deoxyribonucleic Acid
NtC	diNucleotide Conformers
NDB	Nucleic Acid Database
PDB	The Protein Data Bank
PDB (format)	Protein Data Bank (file format)
mmCIF	Macromolecular Crystallographic Information File
CIF	Crystallographic Information File
W-C pairs/pairing	Watson-Crick (canonical) pairing of bases
A	Adenosine-5'-phosphate nucleotide
G	Guanosine-5'-phosphate nucleotide
C	Cytidine-5'-phosphate nucleotide
T	Thymidine-5'-phosphate nucleotide

Table of contents

1 Introduction

One of the most puzzling questions of the past century was how the genetic information is transferred through generations. This question has been gradually answered throughout the years. One of the first researchers addressing the nature of the transfer of genetic material was Frederick Griffith (Griffith, 1928). In 1928, he tried to experimentally prove that bacteria transfer their genetic information by a process called transformation. Avery, MacLeod and McCarty later in 1944 followed in Griffith's footsteps and showed that DNA is the cell's genetic material (Avery, Macleod, & McCarty, 1944). This fact was confirmed in 1952 by Hershey and Chase in their experiment with bacteriophages (Hershey & Chase, 1952). The principle of the preservation of the genetic information was discovered by Watson and Crick in 1953 (Watson & Crick, 1953) when they proposed the now well-known DNA double helical model. The genetic material – nucleic acids, primarily DNA, are one of the most important molecules in molecular biology. Their discovery was a turning point in biology.

Similarly to other biomolecules, nucleic acids are naturally surrounded by water. Water molecules are tightly bound to NAs and create several hydration layers around it. The most important layer is the first hydration layer. Water in this layer behaves very differently from loose water molecules in the bulk phase (J. H. Wang, 1955). It has also been shown that the first solvation layer has an impact on the molecular conformation (Langan et al., 1992) and function.

The solvation layers can be viewed and studied using different methods. One of the widely used methods is structural crystallography, which has been a staple approach for individual molecule study for many years. However modern (structure) science prefers to focus on analyzing large amounts of structures and producing averaged results. Structure information, spatial water distribution and other features, for a variety of molecules is available in structure databases, where data is stored in special file formats. The best known primary structure database is the Protein Data Bank (PDB). We also used a database specialized at description of nucleic acid structures, the Nucleic Acid Database (NDB).

PDB as the structure database for biological macromolecules (Berman et al., 2000) was first established in 1971, however it has begun to expand dramatically in the 1980s and so did the number of its users. File format used to store the data in the Protein Data Bank was historically the PDB (Protein Data Bank) format. Nevertheless, it has recently been replaced

by a richer dictionary of the mmCIF (macromolecular Crystallographic Information File) format.

The PDB format provides description and annotation for protein and nucleic acid structures. It is a set of records in a fixed format, where each line consists of 80 columns, the first six characters hold a record name that describes which type of data the line contains (Westbrook & Fitzgerald, 2003). mmCIF is a newer alternative to the PDB format and is derived from the CIF (Crystallographic Information File) format. It consists of paired data item names and values with significantly more information than the PDB format (Bourne et al., 1997).

The Nucleic acid Database is specialized in description of structural features of nucleic acids (Berman et al., 1992), it contains tables of primary and derivative information in the same file formats as PDB.

Using the data available in these structure databases, we can bioinformatically process thousands of molecular structures including the positions of the waters of hydration. Thus we can obtain the averaged spatial distribution of water molecules around the studied biomolecules as a function of its other properties, such as sequence and/or conformation. The nucleic acid hydration is important for its structural integrity and for its interaction with other molecules. Many molecular interactions take place at the positions of the highest hydration densities of the NA molecules. For example the protein-DNA interaction (Woda, Schneider, Patel, Mistry, & Berman, 1998). Therefore it is very beneficial to identify those positions and possibly model corresponding molecular interactions.

2 Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) is a linear polymer of repeating units called nucleotides. Nucleotides in DNA are composed of a phosphate group, a deoxyribose ring and a nitrogenous base. Nucleotides are linked together into a linear polymer by a so-called phosphodiester link in which the phosphate and sugar group form -P-O-C- covalent bonds.

2.1 Nucleotides

The nucleotide bases are planar aromatic heterocycles. We distinguish four different bases in DNA and classify them into two categories, pyrimidines: thymine and cytosine, and purines: adenine and guanine (Neidle, Schneider, & Berman, 2003). The DNA bases are shown in Figure 2-1. Their usual tautomeric forms are *keto* for pyrimidines and *amino* for purines.

The sugar in DNA is a pentose deoxyribose in its ring form. Each base is linked with the sugar by a glycosidic bond. The deoxyribose moieties are linked by the phosphate groups by two so-called phosphodiester bonds. The nucleotide chemical structure is shown in Figure 2-2. The linear chain of nucleotides forms a single-stranded DNA (deoxyribose sugars) or RNA (ribose sugars).

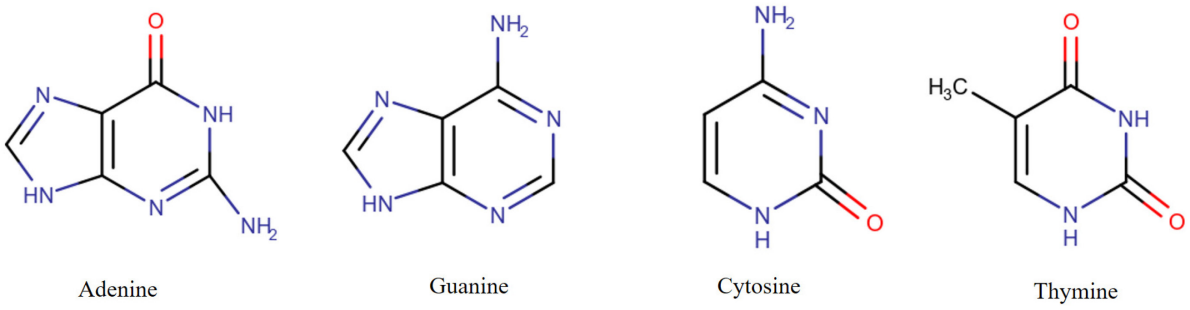


Figure 2-1. The four bases of DNA

2.1.1 Base pairing

Nitrogenous bases can form pairs linked by hydrogen bonds. Although nucleotides can form different types of pairs, the most important type is the Watson-Crick or canonical base pairing shown in Figure 2-2. This pairing is characteristic with its three hydrogen bonds between guanine and cytosine and two hydrogen bonds between adenine and thymine. The arrangement for W-C pairs was first experimentally proven in 1952 by Chargaff (Zamenhof,

Brawerman, & Chargaff, 1952). Watson and Crick then suggested that the exclusive pairing indicates that the DNA structure is double helical (Watson & Crick, 1953). In DNA, the canonical base pairing is strongly preferred so that DNA molecules tend to form double strands with canonical pairs. This ability of self-recognition between two strands is the principal feature which enables the storage of the genetic information.

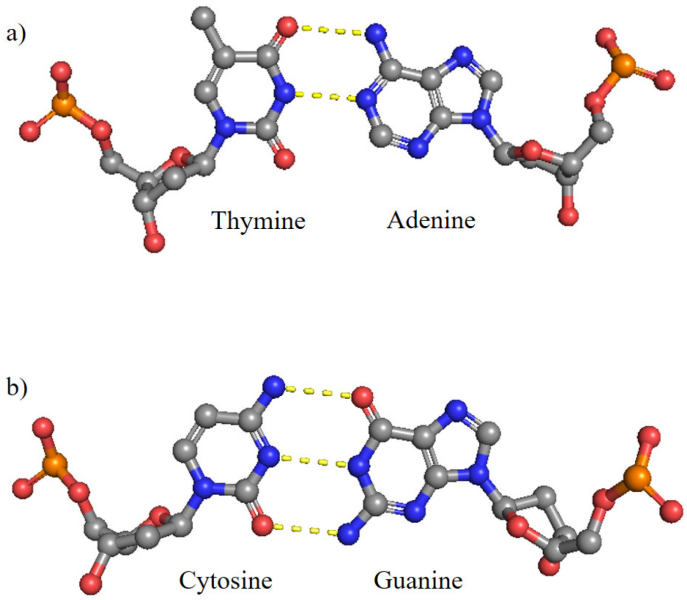


Figure 2-2. Nucleotides forming Watson-Crick (canonical) base pairs. a) Adenine pairs with thymine and b) guanine pairs with cytosine. Pairs were extracted from the Calcium form of B-DNA Undecamer GCGAATTCGCG (Minasov, Tereshko, & Egli, 1999).

W-C pairs have several unique properties which favors them against other pairing types. Firstly, their hydrogen bonds provide high stabilization energies for the molecule. Secondly, both the A-T and G-C pairs in canonical pairing are geometrically similar, having almost the same distance between glycosidic carbon atoms (C1') (Rosenberg, Seeman, Day, & Rich, 1976; Seeman, Rosenberg, Suddath, Kim, & Rich, 1976). And finally, W-C pairing is compatible with the energetically optimal backbone organization in the B-form in DNA or A-form in RNA (Schneider et al., 2018).

2.1.2 Base pair geometry

The nucleotide pairs are formed by hydrogen bonding and therefore show structural flexibility, which is greatly dependent on the neighboring pairs. These structural arrangements can be described by a set of rotational and translational parameters. These are

propeller twist, buckle, inclination and X and Y displacement for one pair, and a rise, helical twist, roll, tilt and slide for two pairs (Neidle et al., 2003). Nevertheless, the details of base pair geometry are not the subject of our study.

2.1.3 Conformations of the sugar phosphate backbone

DNA is a flexible molecule because its phosphodiester backbone has 6 torsion degrees of freedom in each nucleotide building block and the sugar ring is non-planar flexible ring capable of forming different conformations called puckers. A deoxyribose ring has four pseudorotational bonds and one rotational bond.

The pseudorotational bonds in deoxyribose are described by syncyclic torsion angles which have distributions dependent on the chemical nature and conformation of the four substituents of the ring. Rotation of each of the substituents influences the neighbors, so that they are not independent on each other and the ribose pucker can be described by two parameters, so called pseudorotation P and magnitude τ (Altona & Sundaralingam, 1972). There are several sugar pucker geometries, which are distinguished based on the direction of their substituent deviation. The two most common types of deviation in DNA are C2′-*endo*, meaning the C2′ carbon of deoxyribose is deviated from the plane formed by the remaining four sugar ring atoms in the direction of the C4′-C5′ bond and the base. The C3′-*endo* sugar pucker is characterized by the deviation in the same direction but of the C3′ atom. If the deviation is on the opposite side, it is called *exo*.

The six rotational bonds of the phosphodiester backbone are described by the following torsion angles: α, β, γ, δ, ε, and ζ (Figure 3-3).

The spatial distribution of atoms in the phosphodiester backbone is quite restricted, so ranges of their torsion angles are rather limited as well. Distributions of the torsion angles in DNA were studied on a set of structures and presented in 1997 (Schneider, Neidle, & Berman, 1997), but more recently in (Schneider et al., 2018).

The final torsion angle χ describes the rotation around the glycosidic bond between the sugar atom C1′ and the base nitrogen. Due to the spatial arrangement of the other atoms, it can have only two conformational regions: *anti* or *syn*. The *anti* conformation has N1 and C2 atoms of purines and C2 and N3 atoms of pyrimidines oriented away from the sugar ring and has a range between -120° to 180°. The *syn* conformation has, on the other hand, all of these atoms oriented towards the sugar ring and has a range between 0° and 90°.

3 Conformation of DNA

The structure of a DNA duplex was discovered in 1953 (Watson & Crick, 1953). This, today well-known, model of the antiparallel duplex consists of two polynucleotide strands running in opposite directions, wrapped around the central axis. The sugar-phosphate backbone runs on the outside of the molecule. Bases on either of the chains form pairs inside the double helix (Figure 3-1).

DNA is a structurally plastic molecule. Based on its nucleotide sequence, binding partners, or various solute conditions it can form different conformations.

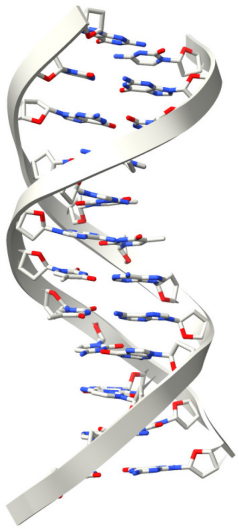


Figure 3-1. The B-DNA dodecamer (Drew et al., 1981).

3.1 Double helical conformations

Depending on the molecular architecture, we distinguish three basic conformation forms (Figure 3-2).

The most important and biologically relevant form is the right handed double helix called B DNA proposed by Watson and Crick. It was confirmed at the atomic resolution by Dickerson's team in 1981 (Drew et al., 1981). Its right-handed structure is characterized by C2′-*endo* sugar pucker and base pairs perpendicular to the central axis. It has a helical pitch of approximately 3.4 Å, 10.5 pairs per turn and the canonical W-C base pairs. The glycosidic bond has the *anti* conformation.

The A DNA conformation is also right-handed, but a slightly wider structure than the B form. It is characterized by C3′-*endo* sugar pucker. The glycosidic bond is also in the *anti* orientation as in B DNA.

The first discovered monocrystal structure of a DNA oligonucleotide formed surprisingly a left-handed duplex called Z DNA (A. H. Wang et al., 1979). Unlike the other two, it has the C2′-*endo* sugar pucker in cytidine (pyrimidine) nucleotides and C3′-*endo* sugar pucker in guanidine (purine) nucleotides. In contrast with the other forms mentioned above, it has a higher helical pitch and more pairs per turn. The glycosidic bond conformation varies in cytidine and guanidine: it has the *anti* conformation in cytidine and the *syn* conformation in guanidine.

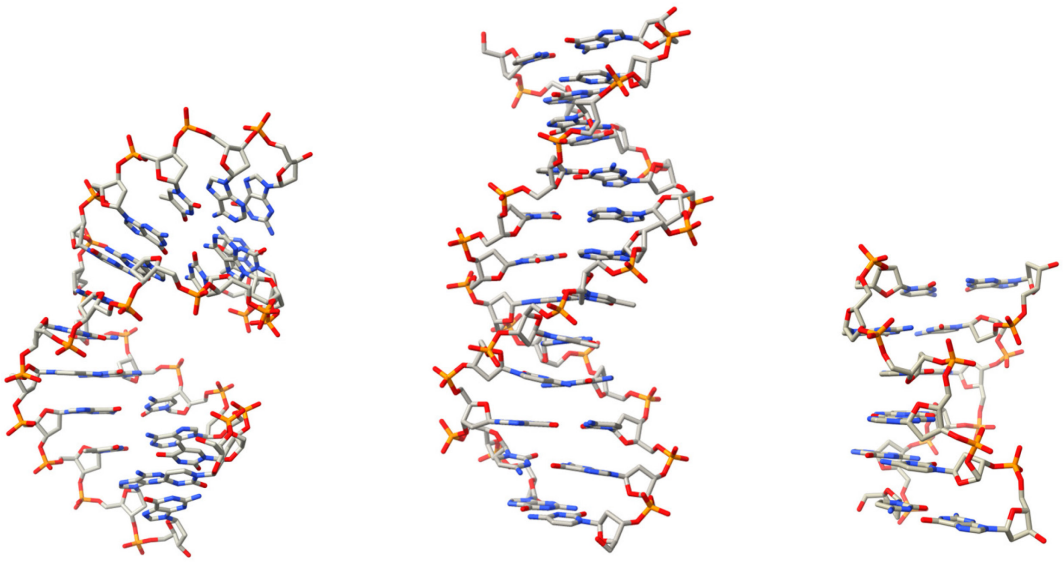


Figure 3-2. The basic confirmation forms. a) A DNA (McCall, Brown, & Kennard, 1985)*, b) B DNA* (Drew et al., 1981)*, c) Z DNA* (Gessner, Frederick, Quigley, Rich, & Wang, 1989).

3.2 Local conformational classes and conformational alphabet

The basic helical conformation forms A, B, and Z describe the global molecular architecture of DNA, but the molecule is locally flexible, especially when binding to other nucleic acids, proteins or different molecules, and their structures cannot be understood and described without going beyond the traditional classification into the A, B and Z forms. DNA conformations therefore need to be described at the local level. The smallest region of a DNA or RNA molecule that can be conformationally examined is a dinucleotide (Richardson

et al., 2008) in which two nucleotides are linked by the phosphodiester link. The region around the central phosphate group carries a large portion of the nucleic acid conformational variability.

Regarding these smaller DNA segments, the polymers can be structurally described by local conformational classes, which are defined by the geometrical properties of the segment. In 2018, a new system of such geometrically defined conformational classes, called NtC classes (diNucleotide Conformer classes), was presented (Schneider et al., 2018). NtC classes were assigned to steps of different DNA structures. Step is a DNA fragment between C5′ of one nucleotide and O3′ of the other in a single strand of a double helical molecule, it contains two deoxyriboses, two bases and a phosphate (Figure 3-3). These fragments were analyzed in a 9-dimensional torsion space defined by 7 backbone torsions and 2 torsions around the glycosidic bond using the weighted k nearest neighbors (k-NN) complete linkage hierarchical clustering with the circular Euclidean distances (Schneider et al., 2018). The paper describes 44 distinct NtC classes that can be assigned at the website dnatco.org. About 20% of the dinucleotides remained unassigned for two reasons: high flexibility of the DNA backbone and the existence of conformationally unique dinucleotides, but mostly because of the fact that especially lower resolution structures are often not refined correctly. The conformity between the analyzed dinucleotide structure and geometries of the NtC classes is measured by a validation score called confal. It is calculated as a Gaussian-weighted distance between the torsion values of the candidate and the average of the NtC class; for definition, see (Schneider et al., 2018).

NtC classes represent a strictly geometric classification of the dinucleotide geometries. A higher level of classification is represented by so-called structural alphabets. They do not consider only strict geometrical properties, but group structural classes, in our case NtC classes, based on some more general features, for example type of stacking. The first introduced structural alphabet for nucleic acids was CANA, Conformational Alphabet of Nucleic Acid (Schneider et al., 2018).

The web service DNATCO (Cerny, Bozikova, & Schneider, 2016) provides assignments of the NtC classes, CANA codes and confal values for all steps in structures containing DNA or RNA available in PDB format version 3.1 or in mmCIF format.

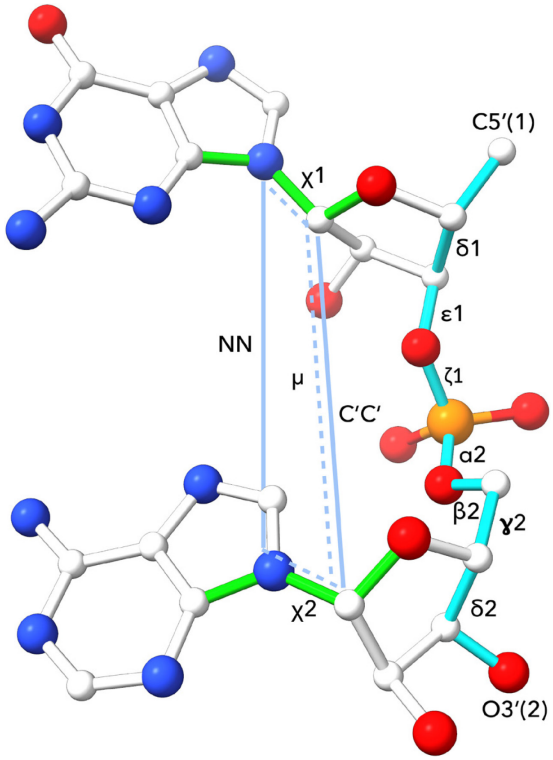


Figure 3-3. Near dinucleotide nucleic acid fragment defining the NtC geometry.

4 The first hydration layer

DNA molecules naturally occur in an aqueous solution and interact with the solvent, which forms solvation layers around the biomolecule. The solvent typical for biomolecules is water, forming hydration layers around the molecule. The earliest analyses of DNA hydration were carried out in the 1950s by gauging the amount of water molecules in the hydration layer (J. H. Wang, 1955). Later on, the exact location of water molecules in the hydration layer was studied, focusing on individual structures that displayed various intriguing water networks (Drew & Dickerson, 1981; Kopka, Fratini, Drew, & Dickerson, 1983). However, the research was then generalized for different structural types and many structures thanks to the data available in nucleic acid databases (Berman, Sowri, Ginell, & Beveridge, 1988; Schneider, Cohen, & Berman, 1992). These later studies have shown that the hydration layers are distinctly ordered.

4.1 The structure of the first hydration layer

From crystallographic studies in the 1980 and later, it had already been known that water molecules in the first hydration layer around the DNA molecules are at least partially ordered (Kopka et al., 1983; Langan et al., 1992). The question remained, if these positions are just idiosyncratic to each newly resolved structure or could be generalized and represent universal properties of the DNA molecule. The first systematic study attempting to answer this question was published in 1992 (Schneider et al., 1992) and followed by a novel way of presenting the water molecule distributions around biomolecular fragments (Schneider et al., 1993). The last cited work presented an analysis of the distribution of water molecules around the DNA bases. Structures of A, B and Z conformations with at least two bases in a strand were searched in a nucleic acid structure database and for each of the bases a hydrated building block was calculated consisting of all the water molecules in a particular distance from the base. Subsequently, water densities for each hydrated building block were calculated using Fourier transform. The positions of the highest water densities in the maps were determined as the most likely positions for water molecules and called hydration sites. The density maps were unique for bases and conformations.

Later study of B DNA molecules showed that the hydration of individual bases and hydration sites are tightly correlated (Schneider & Berman, 1995). Further investigations were focused on the water molecules around phosphate groups (Schneider & Kabelac, 1998;

Schneider, Patel, & Berman, 1998). The former paper (Schneider et al., 1998) also points out the fact that water molecules in dinucleotide steps may overlap depending on the water distribution around phosphates and their mutual orientation.

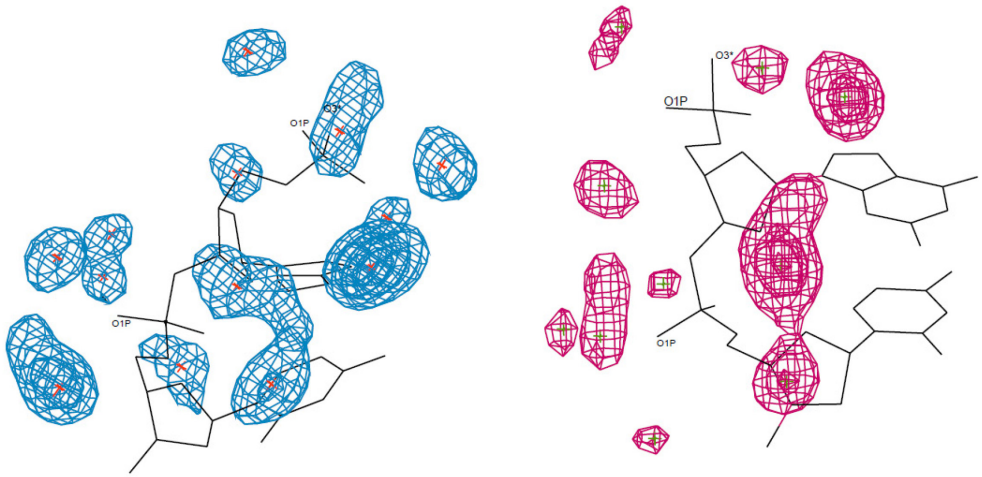


Figure 4-1. Phosphate hydration around the phosphate group in dinucleotides in the so called BI form (left) and in the A form (right) (Schneider et al., 1998).

4.2 The first hydration layer and the NtC classes

As mentioned in chapter 3, NtC classes present a detailed conformational classification for dinucleotides. The first hydration shell is to a large extent organized into well-ordered water positions, so called Hydration Sites (Schneider & Berman, 1995). Since it has been proven that DNA hydration varies depending on the DNA conformation, a question arose whether the hydration of the NtC dinucleotide conformational classes would be distinguishable too. There is currently a project in preparation regarding this issue (Biedermannova, in preparation). The approach to studying the DNA hydration is very similar to the one used earlier, when studying the hydration of amino acids in proteins (Biedermannova & Schneider, 2015) and uses the original method of Fourier averaging (Schneider et al., 1993). The water molecules were divided into two groups, which separately represent the water molecules surrounding the bases and the water molecules surrounding the phosphates and sugars. These two groups of water molecules were selected based on the distance between the water molecule and the particular part of the nucleotide (closer than 3.4 Å).

I will later refer to these groups as the base water and phosphate water. Each of these groups have different properties (Schneider & Berman, 1995; Schneider et al., 1998) and need to be treated separately: hydration around the bases is more concentrated than that around the phosphate groups, which are in double helical DNA open to the solution, and the distributions of water molecules around them are more diffuse.

Classification of the dinucleotide building blocks into the NtC classes allows analysis of hydration of these larger blocks of DNA. In the current Biedermannova study, 3,284 DNA-containing crystal structures were queried from the PDB. All 57,634 analyzed dinucleotide fragments were associated with the crystallographically observed (ordered) water molecules. Distances between the dinucleotides and water molecules were measured and water densities were calculated using Fourier transform. The results are planned to be presented in an atlas of biomolecular hydration, similar to the recently published atlas of the amino acid hydration (Cerny, Schneider, & Biedermannova, 2017). An example of a step with displayed hydration sites around a dinucleotide in the NtC class BB00 (conformation of B DNA) are shown in Figure 4-2. Figure 4-3 shows hydration densities for a few more dinucleotide sequences of different NtC classes and their various hydration sites.

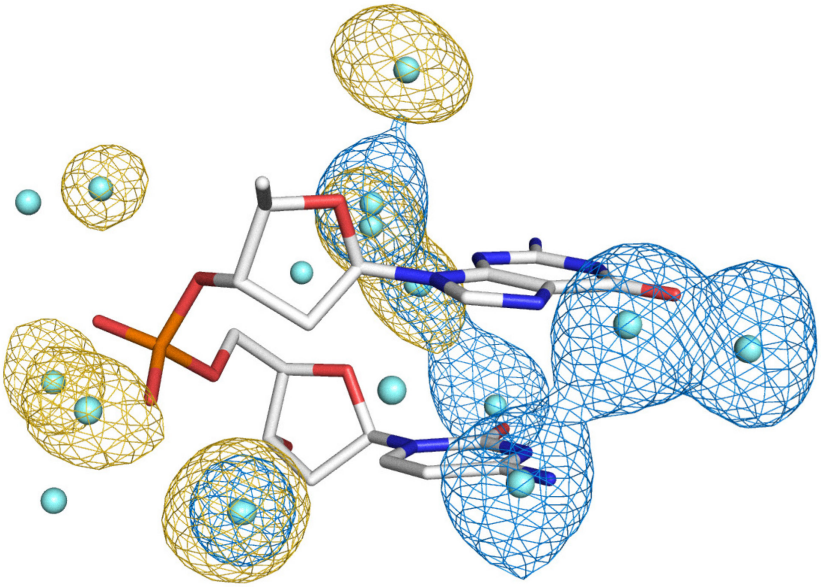


Figure 4-2. The GC step of BB00 NtC class. Blue maps represent hydration sites of base waters, golden maps represent hydration sites of phosphate waters. The density centres of hydration sites are shown as turquoise blue balls.

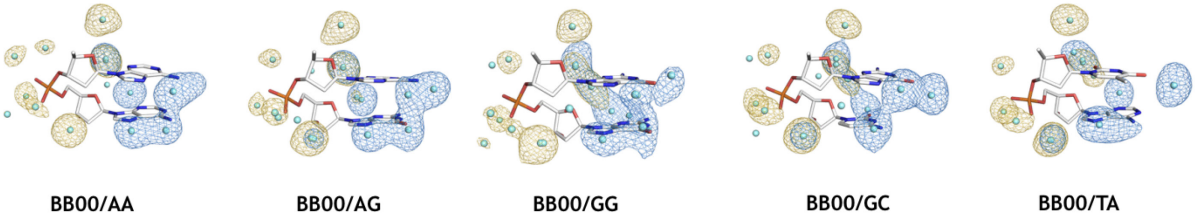


Figure 2. The Hydration Sites around the most common B-DNA form, ***NtC*** class *BB00* in five dinucleotide sequences.

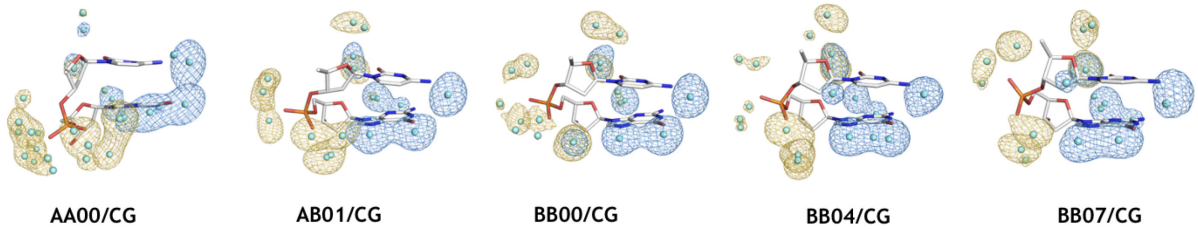


Figure 3. The comparison of the Hydration Sites of dinucleotides with the sequence CG in the ***NtC*** classes *AA00*, *AB01*, *BB00*, *BB04*, and *BB07*.

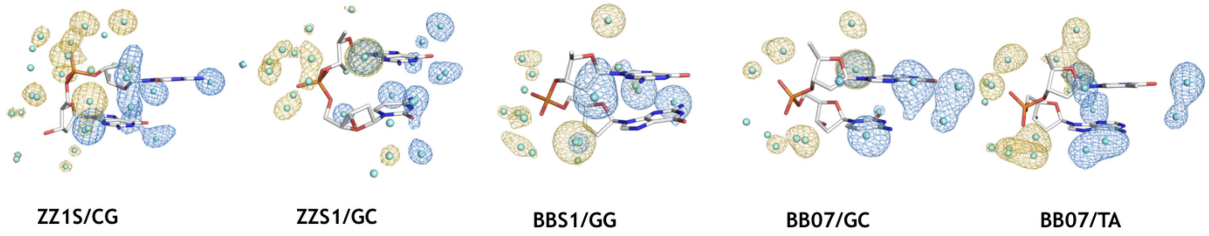


Figure 4. The Hydration Sites around some of the less common ***NtC***.: Z-DNA steps *ZZ1S* and *ZZS1*, *BBS1* observed in quadruplexes, and BII-DNA *BB07*.

Figure 4-3. The hydration sites around a few of the NtC classes.

4.3 The first hydration layer and protein-DNA interactions

The previously mentioned studies showed that both base and phosphate hydration sites are dependent on the sequence and conformation of the DNA molecule. Moreover, it has been confirmed (Woda et al., 1998) that the original proposition by Seeman et al. (Seeman, Rosenberg, & Rich, 1976) that protein atoms involved in binding to DNA take up the positions occupied by solvation water molecules in unbound DNA. Woda et al. calculated protein-DNA intermolecular distances for several protein-DNA complexes and identified potential hydrogen bonds between atoms of the protein amino acid residues and the DNA bases. Distances between the predicted hydration sites and the protein atoms observed in individual structures were measured. The positions of the protein atoms directly bound to the DNA bases were found within 1.5 Å of the predicted hydration positions in 86% of the

interactions based on the independent and previously obtained water densities (Schneider & Berman, 1995).

5 Project

In this project, we investigate the first hydration layer around tetranucleotides built of dinucleotides in the conformations defined by the NtC classes. To obtain a view of the hydration around more extended segments of DNA, we chose to model the hydration sites for canonically paired tetranucleotides as they are observed in the solved crystal structures. These segments are composed of two tetranucleotide strands with all bases forming canonical (W-C) base pairs. Each strand consists of a fragment of four nucleotides starting from a phosphate atom of the first nucleotide to O3′ atom of the fourth nucleotide.

The phosphate linking two central nucleotides of the tetranucleotide fragment is the main focus when studying the conformational and solvation variability. The fragment where we can reliably discuss details of the hydration pattern of the tetranucleotide fragment is highlighted in Figure 5-1 by the dark red box in scheme a) encompassing the central dinucleotide. The hydration pattern of the flanking regions is biased by the fact that nucleotides beyond the tetranucleotide are missing from our model.

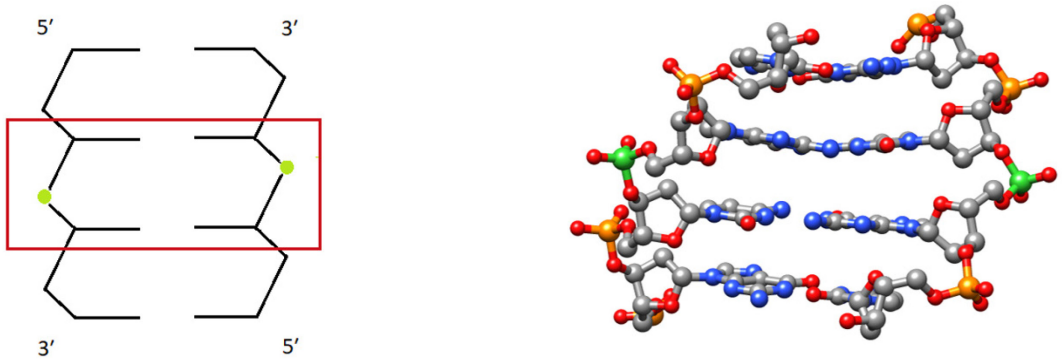


Figure 5-1. The tetranucleotide with the central phosphate highlighted in green. a) Simplified scheme with the central dinucleotides in the red box, b) the tetranucleotide fragment from a solved crystal structure (Garcia-Diaz, Bebenek, Krahn, Kunkel, & Pedersen, 2005).

5.1 Results

DNA-containing crystal structures were queried from the RCSB PDB website and additional information about the steps from the DNATCO website (Cerny et al., 2016). The final obtained data set contained only 27 natural tetranucleotides from 71,313 dinucleotides

from DNA-protein complexes and 7,348 dinucleotides from structures of naked DNA (DNA complexed with no other than small solute molecules).

The original data set of 78,661 dinucleotides structures observed in the database shrank a lot, because only tetranucleotides from structures with the satisfactory parameters were kept: the resolution better or equal to 2.7 Å, the confal value for the overall structure better or equal to 60, the whole tetranucleotide composed of the natural nucleotides, and all four bases forming the canonical (W-C) pairs. Furthermore, to facilitate the analysis for the purpose of this bachelor's thesis, we have decided to take into account only tetranucleotides of the most populous and structurally most important NtC classes: AA00 representing the canonical A form, BB00 representing the most important DNA conformer, BI form, and BB07, the NtC class describing the geometry of the biologically important BII form. In case that the central step of one strand is in the BB07 conformation, the opposite step is classified as the BB00 class and the rest of the steps in the tetranucleotide are classified as either BB00 or BB07. This is the reflection of the fact that two subsequent dinucleotides are never classified as the BB07 class and additionally the dinucleotide of the BB07 class very rarely pairs to a dinucleotide of the same conformation.

Because the number of tetranucleotides fulfilling the above conditions (resolution ≤ 2.7 Å, confal ≥ 60, W-C pairing, unmodified bases and correct NtC classes) was low (27), we were not able to analyze all of the possible tetranucleotide sequences. Some sequences appear rarely or not at all in the analyzed conformational states.

The final dataset of 27 tetranucleotides is summarized in Figure 5-3 regarding the central canonically paired dinucleotides and their corresponding NtC classes. Some of the cells in Figure 5-3 are colored green, which means that the particular hydrated central dinucleotides available from the prior study (Biedermannova, in preparation), which are later used to hydrate the tetranucleotides, have each 1,000 or more water molecules surrounding them. Some dinucleotide sequences of different NtC classes do not have enough water molecules surrounding them (mainly because an insufficient number of relevant structures was available during the previous study), hence are not reliable candidates for the hydration layer analysis. Only the dinucleotide sequences with 1,000 or more water molecules were relevant. Figure 5-2 shows a tetranucleotide scheme with marked central dinucleotides and their NtC classes. This scheme should also explain the notation of the sequences of the central dinucleotides in Figure 5-3.

A lot of the structures in Figure 5-3 were classified as the largest BB00 NtC class and the least number of them belonged to the AA00 NtC class. The ‘BB07 - BB00’ column refers

to the structures which had the central dinucleotide in one strand of the BB07 NtC class and the opposite canonically paired dinucleotide of the BB00 NtC class.

For the final analysis three representatives with the mutually most similar sequences and sufficient water molecules from each of the mentioned NtC groups were chosen and analyzed. These three model structures are written in red in Figure 5-3.

(Note: A dash between two nucleotides (G - C) implies canonical pairing. Similarly, the dash between NtC classes in Figure 5-3 implies canonically paired dinucleotides of the NtC classes. However two nucleotides joined in a word (GC) imply a strand sequence.)

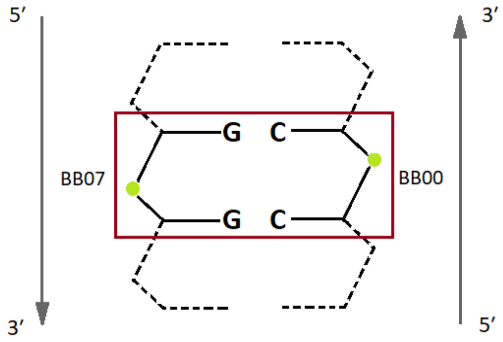


Figure 5-2. The tetranucleotide scheme with the central GG sequence in BB07 conformational class paired with the CC dinucleotide in BB00 class. The NtC classes of the dinucleotides above and below the central one are indicated but they were not constrained to any particular values in our database search.

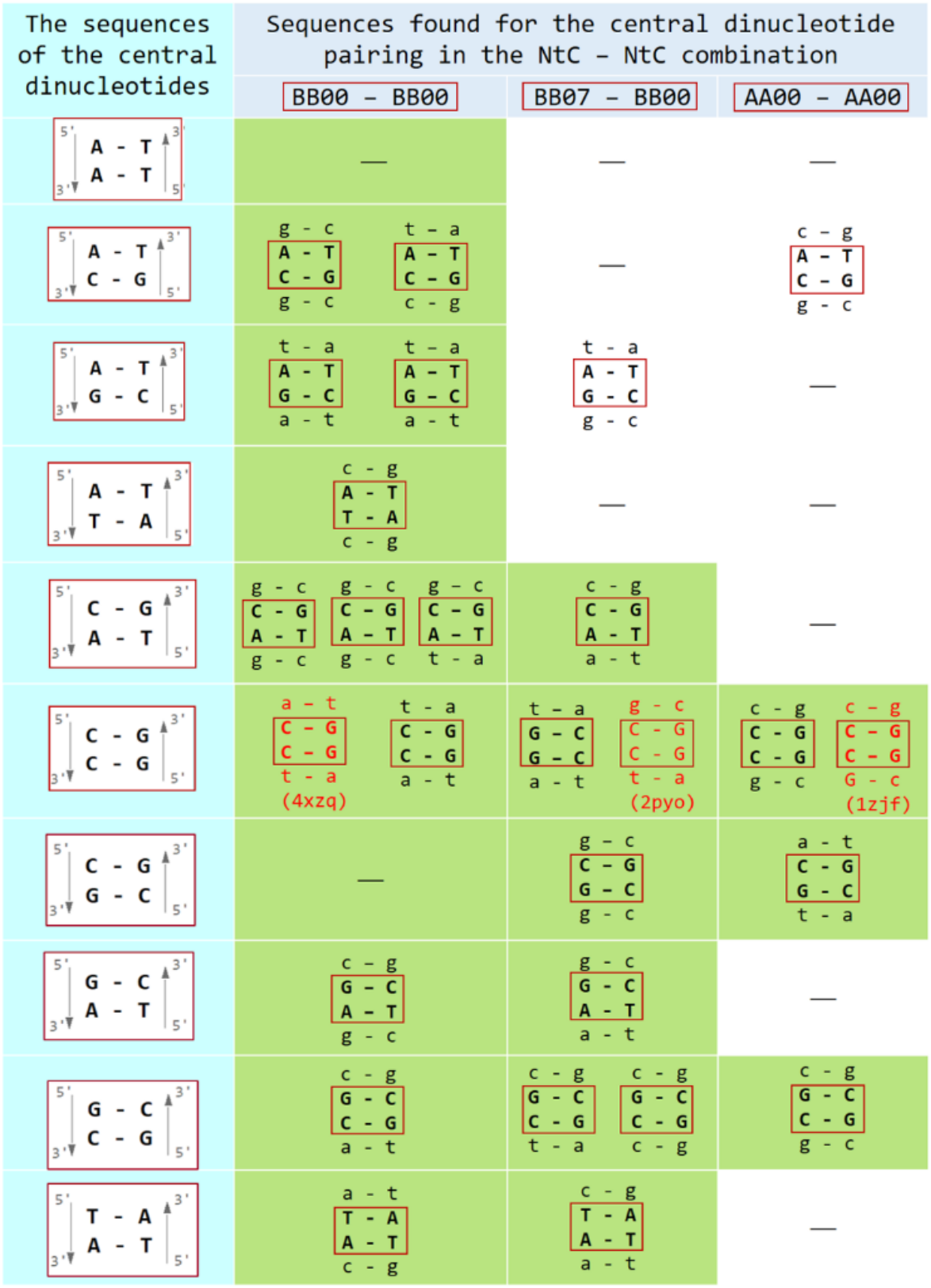


Figure 5-3. Table shows sequences of the 27 tetranucleotides of the final dataset fulfilling these criteria: the central paired dinucleotide is classified in the NtC - NtC combination as specified in the table heading, either BB00 - BB00, BB07 - BB00, or AA00 - AA00. Further, the structure resolution is better than 2.7 Å, all tetranucleotides form the canonical base pairs, and the overall confal value of the structure is better than 60. The three chosen model structures are written in red and their PDB IDs are in brackets. Green color indicates that each of the central canonically paired dinucleotides is hydrated by more than 1,000 water molecules (Biedermannova, in preparation).

The three selected tetranucleotides were hydrated first by the building block of a particular sequence and NtC class containing water molecules connected to the bases and then by the building block containing water molecules connected to the phosphates and sugars. The building blocks with base water molecules and with phosphate water molecules were superimposed onto the tetranucleotide using least-squares fitting as implemented in the molecular modelling system UCSF Chimera (Pettersen et al., 2004). The tetranucleotide with phosphate is shown in Figure 5-4a, and with base water molecules in Figure 5-4b. After adequate file processing, we were left with two files for each tetranucleotide, which included only the fitted base and phosphate water molecules separately.

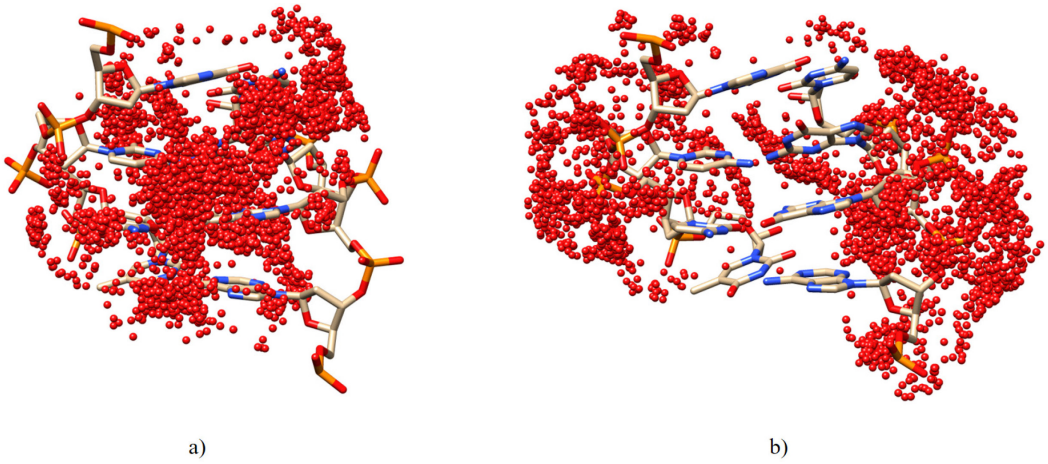
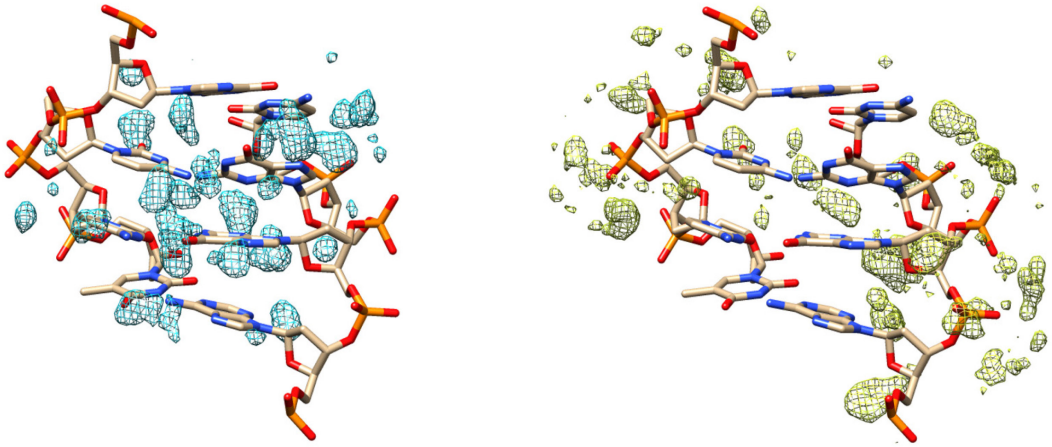


Figure 5-4. The tetranucleotide structure with water molecules from the hydrated dinucleotides superimposed onto them. a) Water molecules around the bases. b) Water molecules around the phosphates and sugars. (Clapier et al., 2008).

Maps of water densities were calculated using the Chimera software from two files containing superimposed base and phosphate water molecules. As shown in Figure 5-5, two density maps, one for the base water molecules and the other for the phosphate water molecules were calculated for each tetranucleotide.



a) b)

Figure 5-5. The tetranucleotide structure with water density maps. a) Base water density map. b) Phosphate water density map. (Clapier et al., 2008).

The resulting water density maps for the three chosen model structures are shown in Figure 5-6, Figure 5-7 and Figure 5-8. The discussion of the results of the first hydration layer around tetranucleotides with canonical pairing will be in detail discussed in the prepared manuscript (Biedermannova, in preparation).

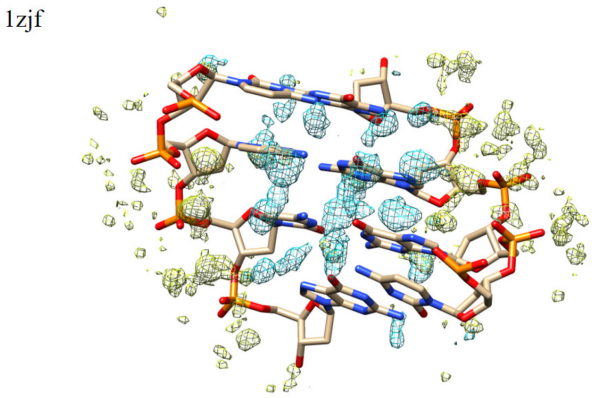


Figure 5-6. The tetranucleotide from the structure of PDB ID 1zjf with the canonically paired dinucleotides in AA00 NtC class and the water density maps. Base water density maps are blue. Phosphate water density maps are light yellow.(Dohm et al., 2005).

2pyo

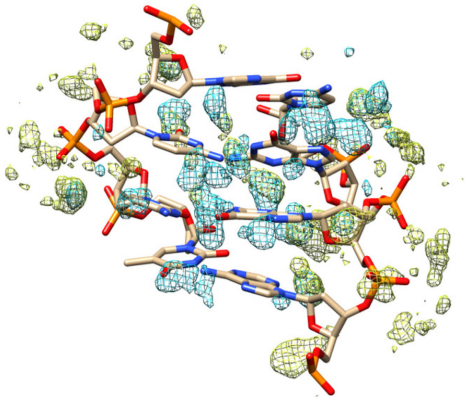


Figure 5-7. The tetranucleotide from the structure of PDB ID 2pyo with the canonically paired dinucleotides in BB00 NtC class on the parallel strand and BB07 on the antiparallel strand and the water density maps. Base water density maps are blue. Phosphate water density maps are light yellow. (Clapier et al., 2008).

4xzq

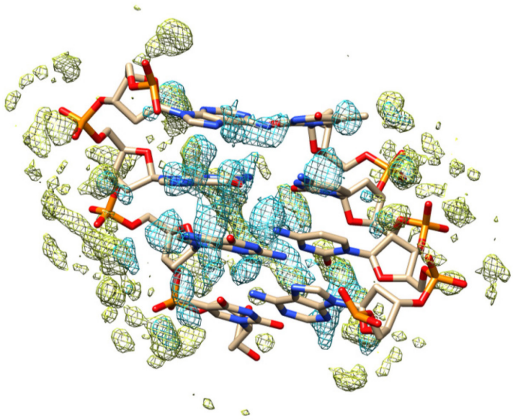


Figure 5-8. The tetranucleotide from the structure of PDB ID 4xzq with the canonically paired dinucleotides in BB00 NtC class and the water density maps. Base water density maps are blue. Phosphate water density maps are light yellow. (*Chatterjee* et al., 2015).

6 Conclusion

This thesis sums up the structural features of DNA molecules and concentrates especially on the discussion of their hydration layers. The DNA molecules are naturally surrounded by water, which bonds to them. These waters of hydration are grouped into several layers and physical properties of these layers are different from the properties of the bulk water. Water molecules especially in the first hydration layer are significantly ordered and form distinct hydration sites. The approaches in studying the hydration of deoxyribonucleic acid have varied over the years concluded by the observation that the hydration structure depends on the DNA conformation and sequence.

The relation between the local dinucleotide DNA conformational classes called NtC (Schneider et al., 2018) and the first hydration layer has been recently investigated using the data available in structural databases and is going to be discussed in a follow up study (Biedermannova, in preparation). Thousands of structures of DNA were bioinformatically processed and averaged water densities for the NtC classes of dinucleotide steps were calculated. The local conformational classes turned out to modify the distributions of water molecules around the DNA and therefore on the interactions with potential binding molecules as well.

In this work, we contributed to the understanding of hydration of DNA by analysis of the hydration patterns in the tetranucleotide double helical fragments in which all bases formed the canonical, Watson-Crick pairs. We limited ourselves to discussion of hydration of three structurally and biologically important conformational classes, BB00 describing the most stable B DNA form, AA00 describing the A form, and BB07 describing biologically important BII form. Additional investigation of perhaps longer than tetranucleotide segments of various sequences, non-canonical pairing or other than the most common NtC conformational classes is needed to fully understand differences in the behavior of water molecules around DNA. As this study demonstrates, the content of the structural databases is rather limited and the suggested studies of longer DNA fragments would require a combination of knowledge-based and modeling approaches.

7 References

Altona, C., & Sundaralingam, M. (1972). Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *Journal of the American Chemical Society, 94*(23), 8205-8212. doi:10.1021/ja00778a043

Avery, O. T., Macleod, C. M., & McCarty, M. (1944). Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med, 79*(2), 137-158. doi:10.1084/jem.79.2.137

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., . . . Schneider, B. (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical Journal, 63*(3), 751-759. doi:10.1016/S0006-3495(92)81649-1

Berman, H. M., Sowri, A., Ginell, S., & Beveridge, D. (1988). A systematic study of patterns of hydration in nucleic acids:(I) guanine and cytosine. *J Biomol Struct Dyn, 5*(5), 1101-1110. doi:10.1080/07391102.1988.10506451

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*(1), 235-242. doi:10.1093/nar/28.1.235

Biedermannova, L., & Schneider, B. (2015). Structure of the ordered hydration of *amino* acids in proteins: analysis of crystal structures. *Acta Crystallogr D Biol Crystallogr, 71*(Pt 11), 2192-2202. doi:10.1107/S1399004715015679

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D., & Fitzgerald, P. M. (1997). Macromolecular Crystallographic Information File. *Methods Enzymol, 277*, 571-590. doi:10.1016/s0076-6879(97)77032-0

Cerny, J., Bozikova, P., & Schneider, B. (2016). DNATCO: assignment of DNA conformers at dnatco.org. *Nucleic Acids Res, 44*(W1), W284-287. doi:10.1093/nar/gkw381

Cerny, J., Schneider, B., & Biedermannova, L. (2017). WatAA: Atlas of Protein Hydration. Exploring synergies between data mining and ab initio calculations. *Phys Chem Chem Phys, 19*(26), 17094-17102. doi:10.1039/c7cp00187h

Chatterjee, N., North, J. A., Dechassa, M. L., Manohar, M., Prasad, R., Luger, K., . . . Bartholomew, B. (2015). Histone Acetylation near the Nucleosome Dyad Axis Enhances Nucleosome Disassembly by RSC and SWI/SNF. *Mol Cell Biol, 35*(23), 4083-4092. doi:10.1128/MCB.00441-15

Clapier, C. R., Chakravarthy, S., Petosa, C., Fernandez-Tornero, C., Luger, K., & Muller, C. W. (2008). Structure of the Drosophila nucleosome core particle highlights evolutionary constraints on the H2A-H2B histone dimer. *Proteins, 71*(1), 1-7. doi:10.1002/prot.21720

Dohm, J. A., Hsu, M. H., Hwu, J. R., Huang, R. C., Moudrianakis, E. N., Lattman, E. E., & Gittis, A. G. (2005). Influence of ions, hydration, and the transcriptional inhibitor P4N on the conformations of the Sp1 binding site. *J Mol Biol, 349*(4), 731-744. doi:10.1016/j.jmb.2005.04.001

Drew, H. R., & Dickerson, R. E. (1981). Structure of a B-DNA dodecamer. III. Geometry of hydration. *J Mol Biol, 151*(3), 535-556. doi:10.1016/0022-2836(81)90009-7

Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., & Dickerson, R. E. (1981). Structure of a B-DNA dodecamer: conformation and dynamics. *Proc Natl Acad Sci U S A, 78*(4), 2179-2183. doi:10.1073/pnas.78.4.2179

Garcia-Diaz, M., Bebenek, K., Krahn, J. M., Kunkel, T. A., & Pedersen, L. C. (2005). A closed conformation for the Pol lambda catalytic cycle. *Nat Struct Mol Biol, 12*(1), 97-98. doi:10.1038/nsmb876

Gessner, R. V., Frederick, C. A., Quigley, G. J., Rich, A., & Wang, A. H. (1989). The molecular structure of the left-handed Z-DNA double helix at 1.0-A atomic resolution. Geometry, conformation, and ionic interactions of d(CGCGCG). *J Biol Chem, 264*(14), 7921-7935. doi:10.2210/pdb1dcg/pdb

Griffith, F. (1928). The Significance of Pneumococcal Types. *J Hyg (Lond), 27*(2), 113-159. doi:10.1017/s0022172400031879

Hershey, A. D., & Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol, 36*(1), 39-56. doi:10.1085/jgp.36.1.39

Kopka, M. L., Fratini, A. V., Drew, H. R., & Dickerson, R. E. (1983). Ordered water structure around a B-DNA dodecamer. A quantitative study. *J Mol Biol, 163*(1), 129-146. doi:10.1016/0022-2836(83)90033-5

Langan, P., Forsyth, V. T., Mahendrasingam, A., Pigram, W. J., Mason, S. A., & Fuller, W. (1992). A high angle neutron fibre diffraction study of the hydration of the A conformation of the DNA double helix. *J Biomol Struct Dyn, 10*(3), 489-503. doi:10.1080/07391102.1992.10508664

McCall, M., Brown, T., & Kennard, O. (1985). The crystal structure of d(G-G-G-G-C-C-C-C). A model for poly(dG).poly(dC). *J Mol Biol, 183*(3), 385-396. doi:10.1016/0022-2836(85)90009-9

Minasov, G., Tereshko, V., & Egli, M. (1999). Atomic-resolution crystal structures of B-DNA reveal specific influences of divalent metal ions on conformation and packing. *J Mol Biol, 291*(1), 83-99. doi:10.1006/jmbi.1999.2934

Neidle, S., Schneider, B., & Berman, H. M. (2003). Fundamentals of DNA and RNA structure. *Methods Biochem Anal, 44*, 41-73.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem, 25*(13), 1605-1612. doi:10.1002/jcc.20084

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., . . . Consortium, R. N. A. O. (2008). RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). *RNA, 14*(3), 465-481. doi:10.1261/rna.657708

Rosenberg, J. M., Seeman, N. C., Day, R. O., & Rich, A. (1976). RNA double-helical fragments at atomic resolution. II. The crystal structure of sodium guanylyl-3',5'-cytidine nonahydrate. *J Mol Biol, 104*(1), 145-167. doi:10.1016/0022-2836(76)90006-1

Schneider, B., & Berman, H. M. (1995). Hydration of the DNA bases is local. *Biophysical Journal, 69*(6), 2661-2669. doi:Doi 10.1016/S0006-3495(95)80136-0

Schneider, B., Boaeikova, P., Necasova, I., Cech, P., Svozil, D., & Cerny, J. (2018). A DNA structural alphabet provides new insight into DNA flexibility. *Acta Crystallogr D Struct Biol, 74*(Pt 1), 52-64. doi:10.1107/S2059798318000050

Schneider, B., Cohen, D., & Berman, H. M. (1992). Hydration of DNA bases: analysis of crystallographic data. *Biopolymers, 32*(7), 725-750. doi:10.1002/bip.360320703

Schneider, B., Cohen, D. M., Schleifer, L., Srinivasan, A. R., Olson, W. K., & Berman, H. M. (1993). A Systematic Method for Studying the Spatial-Distribution of Water-Molecules around Nucleic-Acid Bases. *Biophysical Journal, 65*(6), 2291-2303. doi:Doi 10.1016/S0006-3495(93)81306-7

Schneider, B., & Kabelac, M. (1998). Stereochemistry of binding of metal cations and water to a phosphate group. *Journal of the American Chemical Society, 120*(1), 161-165. doi:DOI 10.1021/ja972237+

Schneider, B., Neidle, S., & Berman, H. M. (1997). Conformations of the sugar-phosphate backbone in helical DNA crystal structures. *Biopolymers, 42*(1), 113-124. doi:10.1002/(sici)1097-0282(199707)42:1<113::aid-bip10>3.0.co;2-o

Schneider, B., Patel, K., & Berman, H. M. (1998). Hydration of the phosphate group in double-helical DNA. *Biophysical Journal, 75*(5), 2422-2434. doi:Doi 10.1016/S0006-3495(98)77686-6

Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A, 73*(3), 804-808. doi:10.1073/pnas.73.3.804

Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Kim, J. J., & Rich, A. (1976). RNA double-helical fragments at atomic resolution. I. The crystal and molecular structure of sodium adenylyl-3',5'-uridine hexahydrate. *J Mol Biol, 104*(1), 109-144. doi:10.1016/0022-2836(76)90005-x

Wang, A. H., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., & Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature, 282*(5740), 680-686. doi:10.1038/282680a0

Wang, J. H. (1955). The Hydration of Desoxyribonucleic Acid. *Journal of the American Chemical Society, 77*(2), 258-260. doi:DOI 10.1021/ja01607a002

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature, 171*(4356), 737-738. doi:10.1038/171737a0

Westbrook, J. D., & Fitzgerald, P. M. (2003). The PDB format, mmCIF, and other data formats. *Methods Biochem Anal, 44*, 161-179.

Woda, J., Schneider, B., Patel, K., Mistry, K., & Berman, H. M. (1998). An analysis of the relationship between hydration and protein-DNA interactions. *Biophysical Journal, 75*(5), 2170-2177. doi:Doi 10.1016/S0006-3495(98)77660-X

Zamenhof, S., Brawerman, G., & Chargaff, E. (1952). On the desoxypentose nucleic acids from several microorganisms. *Biochim Biophys Acta, 9*(4), 402-405. doi:10.1016/0006-3002(52)90184-4